ISOM 671: Managing Big Data (FINAL EXAM)

Name
Email

There are 4 numbered questions spread over 3-pages*. Please submit your responses as a single PDF file by uploading it to the course canvas page. *Notes:

- **Please refer to the notes at the end of this document.**
- **The exam is expected to take about 4-6 hours (or longer); please plan your time efficiently.**
- **You have 12 hours to complete this take-home exam. Start early!**
- **There is a late penalty of 25% for every 6-hour delay.**

---

1. (30 points – 2 points each) Answer the following short questions:

1.1. Discuss two advantages of Hadoop over traditional data warehousing.
1.2. Provide three reasons you wouldn't replace a typical relational database with Hive?
1.3. Discuss what combinations of CAP properties are the most important for NoSQL systems?
1.4. What is "horizontal scaling"?
1.5. Describe "data locality" and how it contributes to Hadoop performance.
1.6. What are two key components/concepts of Hadoop?
1.7. Which tool is best suited to import a portion of a relational database daily into HDFS? Briefly discuss any two reasons for doing so.
1.8. On Hive, what type of join behaves more like a "filter" than a "join"? Explain how it is different from the inner and outer join.
1.9. Which three complex data types are supported by Hive?
1.10. Provide three reasons for using complex data types in Hive.
1.11. Discuss what will be the output for the following Hive query: **"SELECT sentences(txt) FROM quotes_table;"** (assume "txt" is a column with quotes in the table "quotes_table")
1.12. Discuss why the wordcount algorithm performs faster on Spark when compared to Hive.
1.13. If Spark processes queries faster, discuss when not to use Spark SQL instead of HiveQL?
1.14. Briefly discuss the difference between "Tumble Window" and a "Sliding Window." Also show the differences in their pySpark implementation.
1.15. Briefly discuss "Watermark" (in Spark streaming) and its application in a context (e.g., tracking temperature in a cheese warehouse).

---

2. (20 points) In this part, analyze a Kaggle dataset on a mobile advertising campaign[1]. You only need to report the commands/scripts needed to complete the task. You do not need to report the output unless explicitly mentioned.

The data set (a small sample of the original) is downloadable from Canvas and contains the following columns:

1. `id`: ad identifier
2. `click`: 0/1 for non-click/click

[1] https://www.kaggle.com/c/avazu-ctr-prediction/data

3. `hour`: format is YYMMDDHH, so 14091123 means 23:00 on Sept. 11, 2014 UTC.
4. `C1` -- anonymized categorical variable
5. `banner_pos`
6. `site_id`
7. `site_domain`
8. `site_category`
9. `app_id`
10. `app_domain`
11. `app_category`
12. `device_id`
13. `device_ip`
14. `device_model`
15. `device_type`
16. `device_conn_type`
17-24. `C14,C15,...,C21` -- anonymized categorical variables

2.1. (2 pt) Download the dataset and load it in RDS. Using Sqoop, import the data in your hdfs folder: /user/Hadoop/ads/ (*Submit SQL query to load the data in RDS and sqoop command to bring data into hdfs*)
2.2. (2 pt) Create a new internal hive table `ads` based on the dataset - choose appropriate data types for the fields. (*Submit SQL query to create a table in Hive*)
2.3. (2 pt) Load data into the Hive table `ads` and show the first five rows from this table. (*Submit SQL query to load data into Hive table*)
2.4. (4 pt) Use HiveQL to show the top 10 app_ids by CTRs (Click Through Rate = # of clicks / # of impressions). Please show `app_id`, # impressions (each entry in this table is an impression), # of clicks (hint: `SUM(click)`), and CTR for each of these apps. Limit to apps with at least 10 impressions because apps with fewer impressions may not have reliable CTRs. (*Submit Hive query*)
2.5. (4 pt) Using Pig, repeat the query listed in 2.4. (*Submit Pig script*)
2.6. (4 pt) Using Spark SQL, repeat the query listed in 2.4 (*Submit pySpark script with Spark SQL query*)
2.7. (2 pt) Briefly discuss the insights (e.g., query time) you have identified from the three queries (Hive, Pig, Spark). (*Submit up to 150-word discussion*)

---

3. (10 points) In this question, you need to load the Nobel prize winners dataset from canvas (json_award.json) in pySpark. You only need to report the commands/scripts needed to complete the task. You do not need to report the output unless explicitly mentioned.

3.1. (1 pt) Download the dataset and load it in a new directory `nobel` in your S3 bucket. (*Submit S3 URI for the file*).
3.2. (2 pt) Using Spark session (not sc) in pySpark, load the dataset in Spark and register it as a temp table. (*Submit pySpark code*)
3.3. (3 pt) Using Spark SQL in pySpark, write a query that identifies individuals who received the Nobel prize two or more times. (*Submit pySpark code*)
3.4. (4 pt) Using Spark SQL in pySpark, write a query that counts the frequency of words used in the motivation text of the data, and shows the top 5 most frequently used words. (*Submit pySpark code*)

---

4. (10 points) In this question, some scenarios are presented that you need to understand and make sense before answering the following short questions. Please make sure you read the articles provided in the footnote to gain a better sense.

4.1. (3+2 pt) Big data provides rich insights to organizations and enables growth opportunities, but we need to be smart about using this big data[2]. Additionally, Neilsen reported challenges in using big data to understand the audience accurately[3]. The two articles suggest that we need big data, but we must be smart about using this data. Consider you are employed at Google as a Data Scientist, and you are tasked to predict the sales of products in the Google Store[4].

4.1.1. Briefly describe a plan to: a) utilize this dataset (train_v2.csv) - what technologies you will use to process and analyze this data, b) understand the customer base - what insights you will look for in this dataset, and c) how will you predict sales (what variables and model will you use).

4.1.2. Briefly discuss the challenges you expect when making customer targeting (advertising) decisions from the above approach.

4.2. (2+3 pt) Adobe lists six ways to improve customer experience[5], and Talend provides three types of datasets necessary for marketers[6].

4.2.1. Briefly describe and design a Hive database schema that contains all three types of data (including data from Q4.1). (*note: You don't need to list all columns/attributes – just all tables and a few columns (e.g., keys) that are necessary for GStore business*)

4.2.2. For ANY three of the six ways to improve customer experience, provide the three OLAP (data warehouse) star schemas that can help your organization improve customer experience and revenue.

NOTES:

For Q1 (easy), limit your answer to a total of 4 pages (double-spaced) with up to 1 additional page for any exhibits (tables, charts, figures).

For Q2 (medium), you should submit code and scripts (copy-paste in the word file), and screenshots where suggested. For results of any SQL statement, show only the first 5 rows (where asked). It is recommended to use an AWS EMR instance for this question.

For Q3 (medium), you should submit code and scripts (copy-paste in the word file), and screenshots where suggested. For results of any SQL statement, show only the first 5 rows (where asked). It is recommended to use an AWS EMR instance for this question.

For Q4 (hard), please limit the answer to 2 pages of double-spaced text with up to 2 additional pages for exhibits (tables, models, etc.). The grade will not be based on the accuracy of the answer but on your critical reasoning and analytical thinking as reflected in writing; 80% of the score will be allocated to your analysis, critical reasoning, and writing, and only 20% of the score will be allocated for the "answer."

Fine-print:
*Easy: less time taken per point (about 2-5 minutes/point) – 30*3 = 90 minutes*
*Medium: More time taken per point (about 6-8 minutes/point) – 30*5 = 150 minutes (accounting for AWS setup)*
*Hard: Most time taken per point (about 9-15 minutes/point) – 10*9 = 90 minutes*

[2] https://www.mckinsey.com/industries/metals-and-mining/our-insights/adopting-a-smart-data-mindset-in-a-world-of-big-data
[3] https://www.nielsen.com/insights/2023/pros-and-cons-of-big-data-in-audience-measurement/
[4] https://www.kaggle.com/competitions/ga-customer-revenue-prediction/data?select=train_v2.csv
[5] https://business.adobe.com/blog/basics/customer-experience-and-big-data
[6] https://www.talend.com/resources/big-data-marketing/